

# How to make money with classification

Joop van Gent & Piek Vossen

For centuries librarians and other information specialists have promoted classification: the use of predefined categories to make information searchable. And it seems such an obvious truth: if you want to find documents back, you have to store them properly, in cupboards, on shelves, in folders, provided with the appropriate labels.

But since the invention of full text search engines, computers have allowed everyone else to keep their information messy, and just leave it hanging around somewhere, where the computer can find and index it. Even the largest and most messy set of information on earth, the Internet, is easily searchable under the fingertips of everybody. Is there any reason left to argue against the majority of web users, who are hooked on Google?

We believe so. In this article we will show the reader that the “Googlians” are right, but the information specialists are even more right. We claim that it is not only true, that classification does make a lot of sense, but we also show how you can use thesauruses and automatic classification to save money, and make money, in a number of different ways.

## The merits of Google

Google is great. We would be the last persons to deny the enormous impact Google has on the development and popularity of the Internet. People tend to react enthusiastically: “hey, this is fantastic, I want to have this kind of search engine for my own database!” Unfortunately, Google is great for the Internet, but will not work on your information. Why not?

In order to understand why, we have to explain the concepts “recall” and “precision”, often referred to with their more popular equivalents “silence” and “noise”. Recall (silence) indicates which percentage of relevant documents you will miss if you type in a query, and precision (noise) indicates which percentage of irrelevant documents you get. A simple example will clarify this. If you are looking to buy new golf clubs, you could try “golf clubs” for a query, but you will probably not be looking for the yearly gatherings of the Volkswagen community. Still, you will get Google results pointing to VW Golf fellowships. This is called “noise”. With the same query you will miss relevant documents, because they contain the term “golf sticks”, instead of “golf clubs”. This is called “silence”. A bad recall means that you miss a lot of relevant results, and a bad precision means that you get loads of irrelevant results.

The quintessence of Google is, that the focus is fully on good precision, because most people don't care very much about good recall. Why bother that you miss 80% of relevant hits, if you already get 10.000 good ones, a number you won't even take the time to browse through?

But how do you get a very good precision? How is it possible to avoid many irrelevant hits in the top 10? Well, this is the big trick that is

called Google! Google uses two steps to get an excellent precision. The first step is to make sure that the database is outrageously big, so that whatever your type in as a query, you will always get a lot of answers, and the engine will hardly ever leave you with empty hands. The second step is to rank these answers in such a way, that the best ones are always in the top 10. After all, if you know that people will never browse deeper than 100 hits, you don't have to bother about the relevance of hits lower on the list than 100. But how does Google get such a good ranking?

The method Google uses is called **Pigeon Ranking**, and it is very simple. If a query “golf clubs” yields 1.6 million results, all these web pages will contain the words “golf clubs”, but the results appearing in the top 10 are the web pages that **other web sites** refer to most. In the old library world this method was used too, and is called **citation indexing**. Librarians have used citation indexes since ages to keep track of the scientific popularity of books and articles. In books and articles however, these references have a complex structure and are difficult to keep track of, but on the Internet it is relatively easy, because they are just the hyperlinks that you can click on.

By indexing as many web pages as possible, and ranking them according to this ranking method, Google succeeds in making almost every searcher on the web happy.

However, this method will not work for your data. First of all, because – most probably - you don't have that many data. With significantly smaller datasets, the results lists will be smaller too, and the irrelevant results, the “noise”, will become visible at the top. Also, more important, users will find absolutely nothing in many

occasions, just because they did not type in the right words. If your company sells golf clubs, and you always name them this way, a visitor who types in “golf sticks” in the search engine on your web site, will get no results, because the wrong words are used.

The second point is that **pigeon ranking** won't work for your data, because these hyperlinks are made by people in your company, and not – like with Google - by the outside world. In most cases your hyperlinks will not match the perspective of the users, and your visitors will be dissatisfied many times.

Third, Google is not a publisher, but a web search engine provider. Whenever commercial ranking conflicts with what the user wants for result, Google will opt for what the user wants. This is illustrated by the following example. If you type in the name of the company the authors of this article work for, “Irion”, as a query in [www.google.nl](http://www.google.nl), you will get [www.irion.nl](http://www.irion.nl) on top of the results list. There is another Dutch company called Irion, which actually even pays Google for their ranking. They sell lift trucks, but they are not found in the top 10. This is because lesser websites have hyperlinks to their web site.

Does this mean that full text search engines like Google are of no use for information brokers on the web? No, we don't think so, they are in fact a good starting point, but the information provider will have to keep one important principle in mind:

***The smaller your dataset is, the more queries will lead to nothing, and the more attention you will have to pay to match a user's request to the relevant answers.***

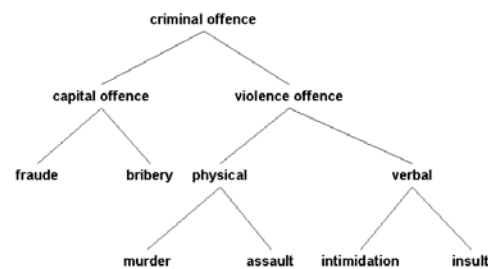
How this can be done, is shown in the next paragraphs.

### Controlled vocabularies

If you want to be sure that user will always find relevant documents in your dataset – provided you can offer them – then you will have to be able to match all kinds of different phrasings of the same information need to one concept. In the abovementioned example of the golf clubs, to which people can refer with the phrase “golf sticks”, a simple **synonyms lexicon** could perhaps do the job: the user types in a query, the system looks it up in the synonyms lexicon, and the original query, together with its synonyms, is offered to the search engine. This approach can give a significant improvement to standard full text search engines.

Even better than synonyms lexicons are semantic networks. Semantic networks are super-vocabularies that contain the common words of speech, and can relate queries to synonyms, hyponyms, hyperonyms or other kinds of semantically related words.

A more common approach is the use of **thesauruses**. Thesauruses are controlled vocabularies with a certain structure, usually designed by library experts for specific domains. They are often dominated by the broader term-narrower term relations.



In a way, they are like semantic networks, but less big and less rich. However, thesauruses are much more refined and tuned to particular domains than semantic networks.

Thesauruses were originally designed to function as an “interlingua”, an intermediate language between the people who offer information, and those who consume it. Documents were simply classified along these controlled vocabularies, and before the invention of full text retrieval technology, libraries, publishers, information brokers **and** end users have always happily worked with them.

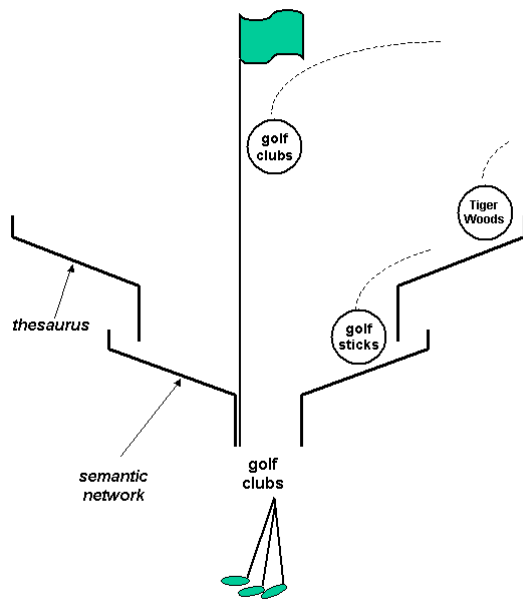
We believe that controlled vocabularies, like semantic networks and thesauruses indeed add much value to full text retrieval solutions, because they can help to match more queries with relevant answers.

To stay with the subject golf, imagine yourself standing on a golf course. How difficult is it to make a hole in one from a long distance, and how easy would it be if you had a big funnel hanging over the hole.

Classification works like a virtual funnel for your query: whatever you type in, the funnel will bring your query to the place where the answers are.

Where under Googlian circumstances only the term “golf clubs” would lead to a web page containing the term “golf clubs”, a semantic network would also match “golf sticks” to the

page, and a thesaurus would even match “Tiger Woods” to the page.



Unfortunately, controlled vocabularies have their serious drawbacks too.

The main drawback with controlled vocabularies in search solutions is, that it is in practice very difficult to make users actually use them. However structured and well designed they might be, they will appear as a complete jungle of words to most users. This is because there is no general agreement about how to structure information on the web, and web site visitors just prefer to skip studying the structure of information for every web site they visit.

Recent research in the publishing industry has shown that only 15% of the professional (!) users is actually willing to use thesauruses as a search method. In the old paper libraries this percentage was significantly higher, but nowadays most people prefer the peculiarities and small discomforts of fast Googlian search over the labor-intensive search through thesauruses.

Second drawback: thesauruses are difficult to develop and maintain. It usually takes many years to develop a good, domain-specific thesaurus, and the development of semantic networks is a monster job. The advantage of most semantic networks when compared to thesauruses is, that they reflect common language use, and can be developed and

maintained independently and relatively easy by many people in different organizations. A well-know worldwide network of people developing a semantic network is the Global WordNet Association. They have been occupied since many years with the development of multi- and cross lingual semantic networks for common language.

Most thesauruses made with blood, sweat and tears are lying on dusty bookshelves, cherished by a small community of people who created and maintained them. This is really a pity, because – as we have illustrated above – thesauruses and semantic networks contribute significantly to the quality of search performance, if used properly.

### To gap the bridge: thesauruses as query catchers

Still, as said, ideally all launched queries should yield a “hole in one”.

Over the last years more and more search engine providers have looked for ways to get the best of both worlds: full text search and search through thesauruses, without having to deal with the disadvantages of each of them. Their solutions in fact always follow the following principle:

***If users don't like to see thesauruses, simply avoid showing them.***

This might seem too simple, but it turns out to be the clue to far more successful retrieval strategies for medium size databases. But how does it work?

Let us get back to our simple example with the golf clubs. We have shown, that if a user types in “golf sticks” instead of “golf clubs”, he/she will miss a lot of relevant information, and in many cases a query will end up with nothing at all, although there *is* actually relevant information in the database. In fact, in the ideal situation a search system should always come up with relevant answers if there are any. This ideal is difficult to achieve, but thesauruses can help a lot in getting closer to the Holy Grail.

Suppose – in the example - there is an underlying, hidden thesaurus, which contains the term “golf equipment”, but none of the terms “golf clubs” or “golf sticks”. Suppose further that all information about golf equipment is gathered and categorized with the thesaurus term ***golf equipment***. Then the only thing a search engine needs to do is match a query like “golf sticks” to the thesaurus term “golf equipment”. In fact the query is ***classified*** in the same way the information in the database

is, and the user does not have to know the structure of the thesaurus to get exactly to the right information.

To put it in a simpler way: if both all documents and all queries can somehow be mapped onto predefined categories, every query will automatically lead to all relevant information.

In many cases there is not just one relevant category, but many. In this case the search engine can present the relevant categories to the user, and let him pick out the correct one. This strategy is called **guided navigation**. Guided navigation is a strategy with which the results of a full text search are classified, and the topmost important categories represented in the result list are shown on top of the results list. In this way the user can pick out the category the actual information request belongs to.

For example: if a user wants to rent an office in Amsterdam, a typical “Googlian” query could be “rent office space”. This will either yield **completely nothing**, or a **large list of results**, containing the words from the query.

In the guided navigation approach the system could – in case of the large list - offer categories like “rent”, “sale”, “background”, “furnishing”, etc., and the user would in this case pick out the category “rent” to cut down the results list to exactly those results that match the information request. This can be done iteratively, and in this way users can easily navigate to enormous amounts of information.

But how can a computer ever **learn** that office space can be rented, sold or furnished? Well, this can be done with machine learning methods. Modern classification systems are based on different pattern matching approaches that can teach the computer to recognize semantic relations between words on the basis of co-occurrence in texts. No specific manual lexicon labor is required here. If a query like “Britney Spears” should lead to “online record shops”, the system can learn this typical domain-specific relationship by simply seeing and analyzing many textual documents.<sup>1</sup>

## Different ways to make money with classification

Search systems are just one area where automatic classification can have its apparent use. As sketched above, they can be improved highly, by classifying both all documents in the database, and all incoming queries. In this way users will not only always get relevant information if there is any, but will also be

guided to the information that the information provider wants to guide them to. In other words: by an intelligent classification of content, information providers can make users end up with exactly the type of information they should get. There is a “Save and Make” money aspect in this. Saving Money, because the classification can be done by an automatic classification system, instead of human beings, and Making Money, because users can easily be guided to those pieces of information that will yield most money, like advertisements.

But this is just the start. In the paragraphs below we describe how automatic classification can be used to save and make money in other areas.

### Personalized information delivery

One obvious and commonly used way to both save and make money with classification is personalized information delivery. This is particularly interesting for information brokers working with online subscriptions to magazines, articles, and news feeds. Customers of these organizations can make profiles for their own fields of interest, by making a selection of categories. In many cases they will pay per field of interest, or they pay per article, and can cut down irrelevant information by selecting fields of interest. In both cases the customers are served better, receive information that matches their personal fields of interest, and the information broker can enlarge his scope of products, and thus his market. The baseline is:

#### **Classify your customers, by classifying your information.**

Examples of Dutch companies that used classification to enable personalized information delivery are Kluwer, MarketXS, Jacobs Company, Bouwradius and Berghauser Pont.

### Personalized advertising

From personalized information delivery it is a small step to personalized marketing / advertising. In current retrieval systems it is quite common to link queries to ads, so that if you type in a query like “golf clubs” you will be confronted with an ad showing shops selling golf equipment. The financial model behind this kind of advertising differs slightly from search engine to search engine, but the basis is that the search engines earn their money through either a pay-per-click-through mechanism, or a pay-per-keyword mechanism. In both cases ads are generated on the basis of a list of predefined keywords that have to be maintained manually.

The problems with the type of query-based ad generation that search engines use is that they will only generate ads for a very small percentage of queries, in practice not more than 10% to 20% of the queries expectedly.

In many European countries agglutinating languages are spoken, languages where compounds are built up by “gluing words together. Even loan words are treated this way: an English query like “hockey team” would be “hockeyteam”, for example in Dutch or Swedish.

If an English query like “hockey stick” leads to a ball sports equipment shop, a query like “hockey shirt” would perhaps do the same because of the word “hockey”, but this would not work for Dutch or German, because in these languages the words would be glued together: “hockeystick” and “hockeyshirt”.

Automatic classification is an excellent way to upturn the business. An experiment with 10.000 random queries entered at the biggest Dutch search engine *Ilse* showed that 81% of the meaningful queries could be lead to a relevant ad by using automatic classification.

### Knowledge Management

In fact, classification means intelligently matching a piece of text of any length or about any topic to one or more categories from a predefined list. As we have seen above, this can be done with queries as well as with whole documents. But if this is the case, it can also be done for any piece of text related to a particular person or expertise. If the thesaurus is not a controlled vocabulary of keywords, but a list of names of experts, the classification system can easily map an information request to the person with the right expertise. And if the thesaurus is a list of “fields of expertise”, the system can lead an information request to a ranked list of expertise’s. This is particularly interesting for organizations working with **helpdesks, call centers, or online help facilities**: requests for information from users/customers will lead to a top 5 of best matching “answers”, no matter whether these answers are in fact expertise’s, persons or just pieces of relevant text containing an answer. There are three levels of knowledge management that can be discerned generally, (1) reveal knowledge, (2) exchange/communicate knowledge, and (3) reason about knowledge. Level (1) is usually implemented in the form of just an intranet where people can publish their knowledge. Level (2) involves intelligent matching of pieces of information that are related. Level (3) usually involves rule sets to enable deduction

of facts from data, generally implemented in the form of artificial intelligence (AI), like decision support systems.

Automatic classification supports level (2) of knowledge management. With a system based on automatic classification, knowledge can be more easily found back, and exchanged, but – more important – pieces of knowledge that belong to each other but are scattered around (either in a database, or in the heads of different people) can be gathered and combined with the help of automatic classification. In the Netherlands we have seen various examples of this type of knowledge management.

A Dutch online insurance company called **Intrasurance** are using automatic classification to interactively build up a knowledge base, that supports their help desk. This is implemented as follows in two steps. Step one: whenever end users approach the help desk with a question, this question is forwarded to an expert, who compiles an E-mail text to cover the answer to the question. Both question and answer are added to an FAQ database, but at the same time they are added to the training set of a classification system. After collection of a large number of question-answer pairs the system is trained, and is now ready for step two. Step two: whenever end users approach the help desk with a question, this question is forwarded to the automatic classification system, which will come up with the most relevant question-answer pairs. The helps desk employee can now pick up the right pair, and cut&edit&paste an answer for the user. The newly constructed answer can be used to train the system or will be forwarded to the relevant expert, who can modify the answer before using it again to train the system. In this easy and interactive way Intrasurance not only builds up a knowledge management system that can gradually take over the expensive and time-consuming tasks of the experts, but also builds up a large knowledge base that makes the company less and less dependent of the knowledge in the heads of experts.

In a third step, the knowledge management system is provided with an interactive shell for en users, so that visitors of the company’s web site, who are not yet customers, can avoid routed phone calls or never-responded E-mails, and will get a much quicker, better and more elaborate answer to any of their questions. This implies a better **conversion rate**: more passers-by turn into customers.

## Filtering and Routing

Apparent uses of automatic classification are filtering and routing. Once a system can discriminate between a number of categories, it can also filter out unwanted information, or route it to certain persons.

A good example of filtering known to us, is one of our own customers, a large multinational that prefers to stay anonymous. They have asked us to develop a system that could automatically filter - from their employees' notebooks - any information that could be judged as insulting or obstructive if examined by the authorities of countries the employees would visit. The actual product developed was an application acting like a virus scanner, but instead of scanning viruses it scanned for potentially insulting or suspicious material in documents.

An example of routing is a pilot we are currently setting up with Dutch local governments. Incoming E-mail with no specific personal E-mail address (e.g. [info@xxxx.com](mailto:info@xxxx.com)) is automatically classified according to a predefined list of governmental policies, and each policy is linked to a list of relevant E-mail addressees, employees within the governmental organization who are responsible for the particular policy. In this way not only human intervention in selecting the E-mails can be avoided, but there is also less risk that the organization misses important mails, because they are not sent directly to the relevant persons.

## Conclusions

The magic combination of thesauruses with automatic classification is in three important ways the key to success for publishers and other information brokers. In the first place it gives a publisher a way to get a good insight in the customer's mind. In the second place it highly improves classic full text search functionality, which will make visitors of web sites and search portals more happy. Third, the technology is the start of a new generation of knowledge management systems, which are far more flexible and less costly than the AI based systems introduced and widely used from the early seventies.

---

<sup>i</sup> "So you want to implement Automatic Categorization?", by R. Kirk Lubbes, CRM, The Information Management Journal, March / April 2003.